Credible Confidence

Albers, Casper J.; Kiers, Henk A. L.; van Ravenzwaaij, Don

Published in:
Collabra: Psychology

DOI:
10.1525/collabra.149

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Albers, C. J., et al. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology,* 4(1): 31. DOI: https://doi.org/10.1525/collabra.149

ORIGINAL RESEARCH REPORT

Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate

Casper J. Albers, Henk A. L. Kiers and Don van Ravenzwaaij

The debate between Bayesians and frequentist statisticians has been going on for decades. Whilst there are fundamental theoretical and philosophical differences between both schools of thought, we argue that in two most common situations the practical differences are negligible when off-the-shelve Bayesian analysis (i.e., using 'objective' priors) is used. We emphasize this reasoning by focusing on interval estimates: confidence intervals and credible intervals. We show that this is the case for the most common empirical situations in the social sciences, the estimation of a proportion of a binomial distribution and the estimation of the mean of a unimodal distribution. Numerical differences between both approaches are small, sometimes even smaller than those between two competing frequentist or two competing Bayesian approaches. We outline the ramifications of this for scientific practice.

Keywords: confidence interval; credible interval; frequentist statistics; Bayesian statistics

The exchange of arguments between frequentist statisticians and Bayesian statisticians goes back many decades. Frequentists rely on the work of classical statisticians such as Fisher, Pearson and Neyman, and apply the lines of thought of these scholars in estimation and inference, most notably in their approach to null hypothesis significance testing (NHST) and the construction of confidence intervals. On the other hand, Bayesians rely on Bayes' paradigm on conditional probability and adjust (subjective) a priori thoughts about the truth – formalized by a probability distribution – into a posteriori statements after observing data.

For many years, the Bayesian approach had two practical disadvantages: (i) many types of models needed a vast amount of computing time, e.g. for estimation through Markov Chain Monte Carlo methods (see, e.g. van Ravenzwaaij, Cassey, & Brown, 2018 for an introduction for psychologists). With the rise of faster computers, this disadvantage has diminished. (ii) Statistical software for researchers within the social sciences, most notably SPSS, as well as teaching of statistical methods relied exclusively on frequentist methods. Nowadays, alternative software with support for Bayesian statistics, most notably R (R Core Team, 2018) and JASP (JASP Team, 2018), are becoming widespread and efforts to teach Bayesian reasoning to social scientists are blossoming (cf. Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2017; Etz & Vandekerckhove, 2018). As a consequence, the Bayesian approach is quickly gaining in popularity.

University of Groningen, NL

Corresponding author: Casper J. Albers (c.j.albers@rug.nl)

The frequentist and Bayesian approaches have fundamental philosophical differences as to how to describe Nature in the form of probability statements. It is obviously important to discuss these differences and the consequences of the choices that both sides make, and this has been done extensively in the (mathematical) statistical literature (cf. Bayarri & Berger, 2004; Pratt, 1965; Rubin, 1984). It is important to have a good, healthy debate between both schools. In general, the criticism of Bayesian methods is that there is too much room for subjectivity (or sometimes not enough, cf. Gelman, 2008), whereas the criticism to frequentist methods is that they are prone to misinterpretation (Bakan, 1966; Cohen, 1994; Goodman, 2008; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Oakes, 1986; Schervish, 1996) and provide answers to unasked questions (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008).

However, too often in our view, the debate is harsh, with Bayesians claiming that all frequentist methods are useless, or vice versa. This style of debating is not new. For instance, over four decades ago Lindley already stated that "the only good statistics is Bayesian statistics" (Lindley, 1975). In recent years, the debate has re-gained popularity due to the increased interest in Bayesian methods in social science research. Furthermore, social media introduced this debate to people previously unaware of this debate. This heated debate has led many non-statisticians to the impression that at least one of the approaches – or, possibly even both approaches – simply must be wrong. An extreme example is that the journal Basic and Applied Social Psychology recently banned frequentist analyses altogether, including reporting of *p*-values, statements

including the word 'significant', etc. (Trafimow & Marks, 2015).

At the core, frequentist and Bayesian approaches have the same goal: proper statistical inference. Philosophical differences in how best to conduct such inference seem less important than the merits of what both approaches have in common. As we will show in this paper, in practice the overlap in uncertainty intervals produced for parameter estimates by both schools is often very large.

Occasionally, the Bayesian and frequentist approach yield substantially different inferences. Usually this occurs when the sample size is very small (see Morey et al. (2016, example 1) and Jaynes and Kempthorne (1976, examples 5 & 6)). It can happen that both approaches yield substantially different outcomes for larger samples, but so far this has only been demonstrated for special 'constructed' examples where, e.g. the space of the outcome variable of interest is highly bi-modal or non-continuous.

Previous work has examined the relationship between the frequentist *p*-value and the Bayesian Bayes factor, both in theory (Benjamin et al., 2017; Johnson, 2005; Marsman & Wagenmakers, 2017) and in practice (Aczel, Palfi, & Szaszi, 2017; Wetzels et al., 2011). In this paper, we examine the similarities between frequentist confidence intervals and Bayesian credible intervals in practice. We will show that in most common cases, the frequentist confidence intervals and Bayesian credible intervals lead to very similar conclusions. By recognizing the near-equivalence between Bayesian and frequentist estimation intervals in 'regular cases', one can benefit from both worlds by incorporating both types of analysis in their study, which will lead to additional insights. We wish to stress that our line of reasoning is not new. For instance, the paper by Bayarri and Berger (2004) starts with "Statisticians should readily use both Bayesian and frequentist ideas. […] The situations we discuss are situations in which it is simply extremely useful for Bayesians to use frequentist methodology or frequentists to use Bayesian methodology". We feel, however, that recent work has stressed differences more than similarities. This paper aims to provide some perspective in this debate.

We shall motivate our opinion on the basis of a series of typical examples from social research. The structure of the paper is as follows. In the next section, we discuss estimation of the population mean in the form of interval estimates. In the section thereafter, we outline, through simulation techniques, the consequences when we are moving away from the 'regular situation' of normally distributed values around a group mean. We end with a discussion including practical recommendations.

Interval estimation of the population proportion

Suppose the interest lies in estimating the proportion of a given population that holds a specific property. This is a very general research question, applicable to many areas: the proportion of diabetes patients that respond positively to a certain treatment method, the proportion of voters expected to vote for a certain political party, the proportion of students passing an exam, etc.

To express the statistical uncertainty about the population proportion, a point estimate alone is not sufficient and an estimate in the form of an interval is preferred. Frequentists call such an interval a confidence interval, Bayesians call it a credible interval. These two types of intervals are, from a theoretical/philosophical point of view, fundamentally different. From a practical point of view, however, both intervals share a common feature: the interval is preferred over the point estimate to express uncertainty. Suppose one estimates a population proportion θ with the interval (.42, .78). This clearly provides different information about the population proportion than the interval (.59, .61), even though in both cases the interval is symmetrical around .60. Furthermore, when a certain value, say .50, is far from the interval, this gives the applied researcher confidence in believing that the unknown true value is unequal to .50: with the interval (.42, .78) one is not keen on rejecting the possibility that $\theta = .50$, whereas with the interval (.59, .61) one can be much more confident about rejecting $\theta = .50$. For this intuitive interpretation, it does not matter whether the interval is constructed using frequentist or Bayesian methods.

There are different frequentist and Bayesian approaches to generating such intervals, all based on a random sample of n objects, of which it is recorded that m objects hold the property of interest. These models differ in the mathematical way they are constructed, yet all are sensible approaches to estimating a proportion. Below, we outline three common frequentist approaches and two common Bayesian approaches. For sake of simplicity, we set the confidence/credible level at a fixed value of 95%. Furthermore, we assume that the population size is much larger than the sample size, such that we do not need to worry about finite population corrections.

Approach F1: Plus four method

When n, np and $n(1-p)$ are all not 'too small', an approximate confidence interval is directly obtained from the normal approximation $\text{Bin}(n, p) \approx \text{N}(np, np(1-p))$ due to the Central Limit Theorem. This gives the interval

$$\hat{p} \pm 1.96\sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)}, \tag{1}$$

where $\hat{p} = m/n$ is the observed proportion in the sample and 1.96 is the percentile of the standard normal distribution corresponding to the 95% level.

This asymptotic approach can be improved upon through the so-called plus-four method (Agresti & Coull, 1998). In this method, the estimate $\hat{p}$ in (1) is replaced, on all three instances, by $\tilde{p} = (m + z)/(n + 2z)$, where $z = 1.96$. Roughly, this method adds two successes and two failures to the sample, hence the name plus-four method. For large samples this change has little effect: the difference between $\tilde{p}$ and $\hat{p}$ is relatively small. For smaller

samples Agresti and Coull have shown that their method constitutes an improvement.

Approach F2: Exact confidence interval

Approach F1 is asymptotic and – even with the "plus four"-correction outlined – does not necessarily work well for small samples. However, it is frequently used, mainly because of its simplicity and the lack of alternative methods available in common software packages. Blyth (1986) discusses a method for computing the exact confidence interval, after Clopper & Pearson (1934):

$$\left(\left(1+\frac{n-m+1}{mA}\right)^{-1},\left(1+\frac{n-m}{(m+1)B}\right)^{-1}\right)$$

with $A = F_{0.025;\,2m,\,2(n-m+1)}$ and $B = F_{0.975;\,2(m+1),\,2(n-m)}$ being percentiles from F-distributions.

Approach F3: through arc sine transformation

This approach is based on the approximation (cf., Shao, 1998) that

$$\mathrm{var}\left(\sin^{-1}\sqrt{p}\right)\approx\frac{1}{4n}$$

which, after some derivations, leads to the interval

$$\left(\sin^2\left(\sin^{-1}\sqrt{\hat{p}}-\frac{z}{2\sqrt{n}}\right),\sin^2\left(\sin^{-1}\sqrt{\hat{p}}+\frac{z}{2\sqrt{n}}\right)\right).$$

One of the instances where this approach is used is in the computation of Cohen's *h*.

Table 1: 95% confidence/credible intervals for the five methods for various settings of *m* and *n*.

Method	$n = 40$, $m = 10$	$n = 40$, $m = 20$	$n = 80$, $m = 40$	$n = 500$, $m = 250$
F1	(.134, .410)	(.345, .655)	(.390, .610)	(.456, .544)
F2	(.127, .412)	(.338, .662)	(.386, .614)	(.455, .545)
F3	(.119, .429)	(.305, .743)	(.357, .667)	(.440, .564)
B1	(.142, .403)	(.351, .649)	(.393, .607)	(.456, .544)
B2	(.136, .398)	(.350, .650)	(.392, .608)	(.456, .544)

Approach B1: uniform prior

Bayesian approaches are specified through their prior distribution. The beta-distribution is a so-called conjugate prior of the Binomial distribution, which means that the posterior distribution is also Beta. In general, when using a Beta(a, b) distribution as prior, the posterior is given by the Beta($a + m$, $b + n - m$) distribution. By taking the 2.5% and 97.5% percentile points of this distribution, one achieves the 95% credible interval.

Approach B1 is based on the prior assertion that all values for p between 0 and 1 are equally likely. This is achieved by using the uniform(0,1) distribution, which is identical to the Beta(1, 1) distribution, as prior. This results in a Beta($1 + m$, $1 + n - m + 1$) as posterior.

Approach B2: Jeffreys prior

Jeffreys prior is a so-called non-informative prior (which means it is invariant under reparametrizations of the problem space), which is a desirable property of a prior. The Jeffreys prior for the current setting is the Beta(½, ½) distribution, yielding the Beta($½ + m$, $½ + n - m$) posterior.

Comparison

Table 1 lists the intervals obtained by the five methods for various choices of *m* and *n*. It is clear that the methods are in general agreement, especially when *n* is large. Only exception is the arcsine method, that consistently provides wider intervals. In **Table 2**, we study the five approaches in more detail. For various choices for *n*, it lists the average overlap between approaches for all possible values of *m* (i.e. $m = 0, 1, \ldots, n$). The arcsine method clearly has different behavior than the four others. For those other methods, even with *n* as low as 10, the overlap between any two approaches, whether one is Bayesian and the other frequentist, or whether both are from the same 'school', is at least 90%. For these methods, the agreement increases if *n* increases. Both Bayesian approaches are usually, but not always, somewhat more similar to each other than to the frequentist approaches, and the same can be said for the frequentist approaches F1 and F2. However, the differences are negligible. Thus, a frequentist might have the same level of agreement with a fellow frequentist as with a Bayesian. Similarly, it is entirely possible that two Bayesians agree less with each other than with a frequentist. In the words of the Bayesians Jaynes and Kempthorne (1976, p. 195): "The differences are so small that I could not magnify them

Table 2: Overlap between methods. Overlap between approaches *A* and *B* is computed as the average of the percentage of the CI of *A* that is also covered by the CI of *B*, and the percentage of the CI of *B* also covered by *A*'s interval.

n	F1–F2	F1–F3	F2–F3	F1–B1	F1–B2	F2–B1	F2–B2	F3–B1	F3–B2	B1–B2
10	.978	.924	.935	.913	.916	.930	.933	.890	.903	.970
25	.978	.893	.909	.950	.948	.950	.947	.873	.885	.971
50	.978	.879	.896	.969	.965	.962	.958	.867	.877	.975
100	.980	.869	.884	.981	.977	.971	.968	.862	.869	.980
500	.987	.852	.861	.994	.989	.985	.984	.850	.853	.990

into the region where common sense is able to judge the issue".

Interval estimation of the population mean

Methods

For continuous data, the central limit theorem states that for any reasonable n, the sampling distribution of the sample mean is approximately normal. A frequentist 95% confidence interval for the population using the commonly used t-distribution is as follows

$$\bar{x} \pm t_{n-1} s / \sqrt{n}, \quad (2)$$

where $\bar{x}$ is the sample mean, t_{n-1} is the corresponding critical value from a t-distribution with $n - 1$ degrees of freedom, s is the sample standard deviation, and n is the sample size. We are going to contrast this standard frequentist confidence interval with a Bayesian credible interval, based on a default Cauchy prior on effect size, as this is currently implemented in e.g. the 'point-and-click' programmes JASP (JASP Team, 2018) and jamovi (jamovi project, 2018). The construction of such an interval proceeds as follows.

A prior is constructed for the population effect size delta, such that $\delta \sim N(0, \sigma_\delta^2)$ and $\sigma_\delta^2 \sim$ Inverse $\chi^2(1)$. Combining these two yields $\delta \sim$ Cauchy (Liang, Paulo, Molina, Clyde, & Berger, 2008). The next step is the construction of a likelihood function: L(data$|\delta$). The posterior is proportional to the product of the prior and the likelihood. The 95% credible interval constitutes the middle 95% of this posterior.

With these restrictions in place, we conducted two sets of simulations. In the first set, we generate normally distributed data for a single group that varied along the following two dimensions:

1. Corresponding t-statistic: 0.5, 1, 1.5, and 2 (i.e., a sample of generated values was transformed such that the corresponding t-values exactly equaled these values, and that the sample standard devation equaled 1);[1]
2. Number of participants: 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, and 30.

Subsequently, we calculated 95% confidence and credible intervals for the resulting data.

In the second set of simulations, the data is artificially constructed such that the data vary on how skewed the underlying population distribution is. This was done by simulating data using the rsn function in R (from package sn, see Azzalini, 2017). Skew was manipulated by varying the 'alpha' parameter from 0 to 10 in steps of 1 (see Azzalini, 2014 for details). The number of participants was fixed to 20 for this set of simulations. Subsequent to sampling from the skewed normal distribution, the data was standardized and $t/\sqrt{20}$ was added to each data point to ensure all simulations varied only along the value of the t-statistics and the alpha parameter. Finally, we calculated 95% confidence and credible intervals for the resulting data.

Results

Results for the first set of simulations, based on normally distributed data, are shown in **Figure 1**. The figure shows that frequentist confidence intervals and Bayesian credible intervals correspond closely. For lower sample sizes, the confidence intervals appear to be marginally wider than the credible intervals, but this difference quickly disappears for more realistic (but still small) sample sizes.[2]

Results for the second set of simulations, based on right-skewed data, are shown in **Figure 2**. The results of this second set of simulations mirror those of the first set of simulations in that there is no qualitative difference between the confidence and credible intervals. This is perhaps not so surprising: although the data itself deviates from normality, the central limit theorem implies that the sampling distribution of the sample mean is still approximately normal. As such, there is no reason to expect substantial differences between both sets of simulations.[3]

Discussion

In the present paper, we have demonstrated by means of various examples that confidence intervals and credible intervals, in various practical situations, are very similar and will lead to the same conclusions for many practical purposes when relatively uninformative priors are used. The examples used here are based on small samples but are otherwise well behaved and could easily occur in practice. When sample size increases, the numerical difference between both types of interval will (usually) decrease.

So in what situations do the approaches yield more substantial differences? There are two main examples: (1) restriction of range of the data; (2) Bayesian methods based on a considerably more informative prior. As an example of the first point, consider 15 scores on a Likert scale ranging from 1 to 5. Suppose that ten scores are 1, four scores are 2, and one score is 5. Construction of a classical 95% confidence interval results in the interval (0.95, 2.12), an interval that includes values below the minimum possible value of 1. The Bayesian 95% credible interval is bounded by definition to not include values beyond the range of the parameter space. For a uniform prior on this interval, combined with the assumption that the sample standard deviation equals the population standard deviation, the resulting 95% credible interval is (1.08, 2.07) (see **Figure 3**).

The second point highlights the scope of our present findings: we have shown numerical similarities between frequentist and Bayesian methods for (relatively) uninformative priors. Depending on the research context, vastly different intervals can be obtained if one chooses a specific informative prior. Our paper meant to highlight similarities when relatively standard, off-the-shelve, methods are used for constructing intervals under both regimes, using 'objective' or fairly uninformative priors, in the simple common contexts of estimation of proportions and means.

Why then, in cases with little or no prior information, bother with Bayesian approaches, and not stick to the

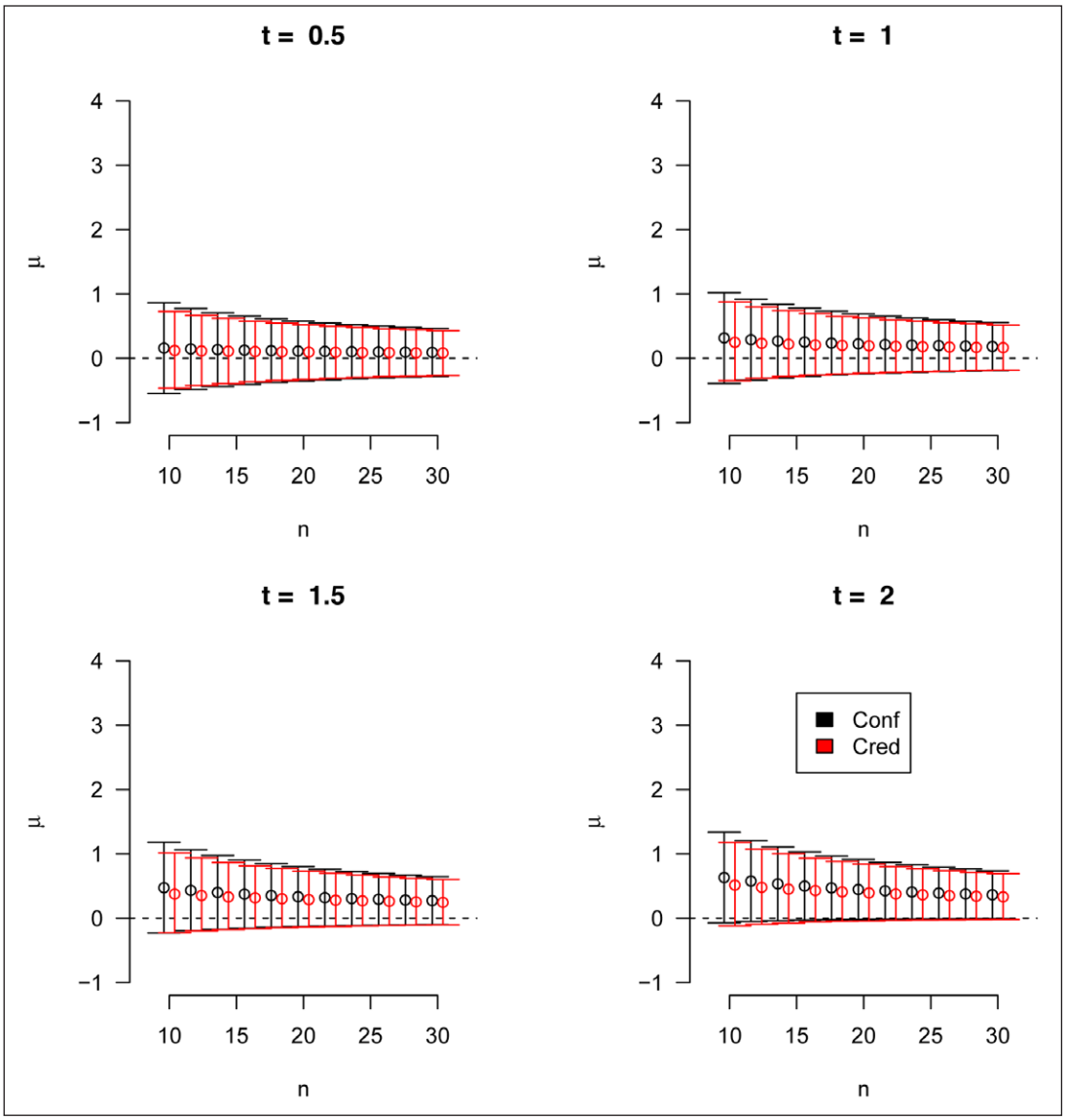

Figure 1: Comparison of 95% confidence intervals (black) to 95% credible intervals, based on the default Cauchy prior (red) for Normally distributed data. Results show intervals are nearly identical.

more traditional frequentist confidence interval? A good reason is that a Bayesian analysis is more in line with the way researchers actually interpret their results (whether frequentist or not). That is, researchers tend to interpret their results in explicit or implicit terminology indicating how certain they are about what the effect size truly (i.e. in the population) is. As many papers and text books emphasize, frequentist approaches cannot warrant such statements, but Bayesian approaches can: One can claim that there is a 95% chance that the true effect size is in the credible interval. Even stronger, one can accompany the credible interval with a *full picture* of the distribution from the true effect size by means of giving the full posterior distribution, see **Figure 3** for an example. Similar frequentist approaches to distributional inference exist (Albers, 2003; Kroese & Schaafsma, 2004), but are neither straightforward nor often used in practice. A frequentist analogue to the rich information provided by the posterior distribution is the bootstrap (Efron & Tibshirani, 1994).

The frequentist approach works from the premise that only the data are prone to random fluctuations, while the true effect is fixed, and hence it makes no sense to specify probabilities for the (fixed) population effect size but only about the probability as to whether the confidence intervals estimated by means of the data will cover the true effect size. This is a subtle difference with the Bayesian credible interval interpretation, but as the way people like to interpret results is more in line with the latter, the Bayesian approach is better in serving researchers at their wishes. This comes with a price, however. The price is that the statements are always conditional upon the prior that one has specified. Fortunately, however, the exact location of credible intervals does not appear to vary strongly with variations in the prior. Indeed, in the case where we assume that the population variance is known, the confidence interval for means can be obtained by a particular choice of the prior, namely the uniform prior. This is implausible in practice, but can be seen as a limiting case of a flat prior. And as we have seen now,

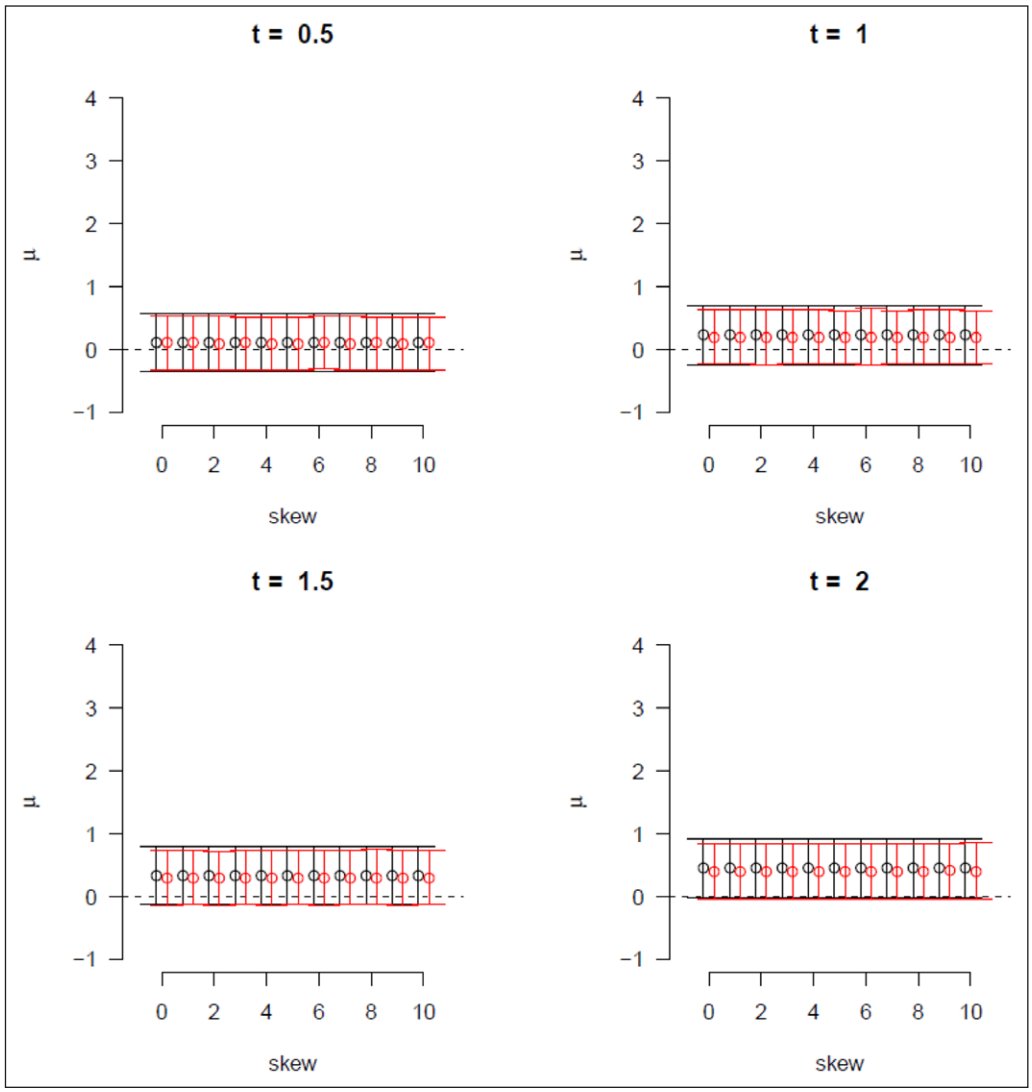

Figure 2: Comparison of 95% confidence intervals (black) to 95% credible intervals, based on the default Cauchy prior (red) for right-skewed data. Results show intervals are nearly identical.

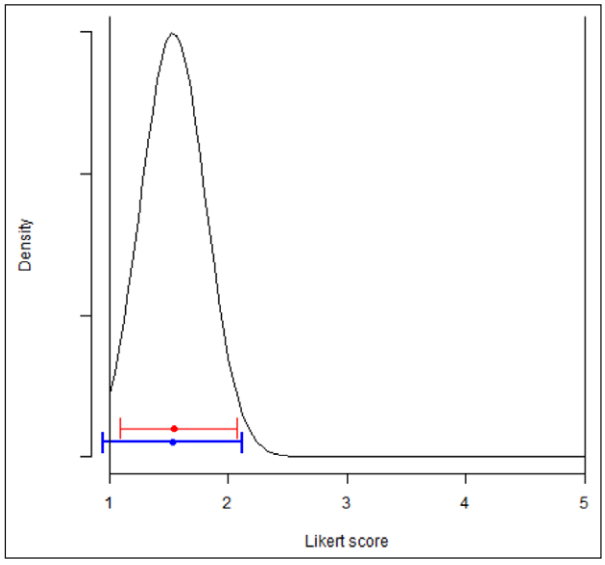

Figure 3: Posterior density, credible interval (red) and confidence interval (blue) for the example with 15 measurements on a Likert-scale.

it does not lead to very different intervals than does the more realistic Cauchy prior.

For us the main message of our paper is as follows. Frequentist confidence intervals can be interpreted as a reasonable approximation to a Bayesian credible interval (with uninformative prior). This is reassuring for those who struggle with the formally correct interpretation of frequentist intervals. Additional insight can be obtained when these intervals are complemented (or replaced) by a full posterior distribution for the effect size measure under study. The posterior distribution will, conditionally upon a chosen prior, give the full picture of the uncertainty around its possible value. It can provide information on skewness, bimodality, and other properties – or the lack thereof, such as in **Figure 3** – that a simple interval, with only a lower and upper bound, can not. Furthermore, it can estimate the probability that the parameter is larger or smaller than a fixed value, e.g. 0 or 0.5, or is within a certain interval. As such, posterior distributions can ideally work towards the enhancement of science.

Data Accessibility Statement

All computations have been performed using *R* (R Core Team, 2018). All software code is available from https://osf.io/dgfht/.

The preprint of this paper has also been published on OSF.

Notes

[1] With this specification for each data set $s = 1$, and the sample mean equals $t/\sqrt{n}$, which implies that the confidence interval in (2) is exactly specified as $t/\sqrt{n} \pm t_{n-1}/\sqrt{n}$.

[2] We replicated this simulation study 25 times, and only found negligible differences in lower and upper bounds, see supplementary materials.

[3] We also replicated this simulation study 25 times and again only found negligible differences in the lower and upper bounds, see supplementary materials.

Competing Interests

The authors have no competing interests to declare.

Author Contribution

CJA conducted the intervals for discrete data analysis and wrote the first version of the manuscript, HALK wrote a part of the paper, and critically commented on various versions of the manuscript, DvR conducted the intervals for continuous data simulations and critically commented on various versions of the manuscript. All authors contributed to writing of the final manuscript.

References

Aczel, B., Palfi, B., & **Szaszi, B.** (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE, 12*(8). DOI: https://doi.org/10.1371/journal.pone.0182651

Agresti, A., & **Coull, B. A.** (1998). Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions. *The American Statistician, 52*(2), 119–126. DOI: https://doi.org/10.2307/2685469

Albers, C. J. (2003). *Distributional inference: The limits of reason*. s.n. Retrieved from: http://hdl.handle.net/11370/3ee16f7c-e7e0-4cb5-9f28-3037eacdfb6d.

Azzalini, A. (2014). *The skew-normal and related families*. Cambridge, UK: Cambridge University Press.

Azzalini, A. (2017). *The R package sn: The Skew-Normal and Related Distributions such as the Skew-t* (version 1.5-1). Università di Padova, Italia. Retrieved from: http://azzalini.stat.unipd.it/SN.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423–437. DOI: https://doi.org/10.1037/h0020412

Bayarri, M. J., & **Berger, J. O.** (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science, 19*(1), 58–80. DOI: https://doi.org/10.1214/088342304000000116

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Johnson, V. E., et al. (2017). Redefine statistical significance. *Nature Human Behaviour, 1*. DOI: https://doi.org/10.1038/s41562-017-0189-z

Blyth, C. R. (1986). Approximate Binomial Confidence Limits. *Journal of the American Statistical Association, 81*(395), 843–855. DOI: https://doi.org/10.2307/2289018

Clopper, C. J., & **Pearson, E. S.** (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika, 26*(4), 404–413. DOI: https://doi.org/10.2307/2331986

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. DOI: https://doi.org/10.1037/0003-066X.49.12.997

Efron, B., & **Tibshirani, R. J.** (1994). *An Introduction to the Bootstrap*. CRC Press.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & **Baribault, B.** (2017). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review,* 1–16. DOI: https://doi.org/10.3758/s13423-017-1317-5

Etz, A., & **Vandekerckhove, J.** (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review, 25*, 5–34. DOI: https://doi.org/10.3758/s13423-017-1262-3

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis, 3*(3), 445–449. DOI: https://doi.org/10.1214/08-BA318

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135–140. DOI: https://doi.org/10.1053/j.seminhematol.2008.04.003

Jamovi project. (2018). *jamovi (version 0.8) [computer software]*. Retrieved from: https://www.jamovi.org.

JASP Team. (2018). *JASP (Version 0.9.0.1) [Computer software]*. Retrieved from: https://jasp-stats.org/.

Jaynes, E. T., & **Kempthorne, O.** (1976). Confidence Intervals vs Bayesian Intervals. In: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, 175–257. Springer, Dordrecht. Retrieved from: https://link.springer.com/chapter/10.1007/978-94-010-1436-6_6. DOI: https://doi.org/10.1007/978-94-010-1436-6_6

Johnson, V. E. (2005). Bayes Factors Based on Test Statistics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67*(5), 689–701. DOI: https://doi.org/10.1111/j.1467-9868.2005.00521.x

Kroese, A. H., & **Schaafsma, W.** (2004). Distributional Inference. In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/0471667196.ess0628.pub2

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & **Berger, J. O.** (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association, 103*(481), 410–423. DOI: https://doi.org/10.1198/016214507000001337

Lindley, D. V. (1975). The Future of Statistics: A Bayesian 21st Century. *Advances in Applied Probability, 7*(supplement), 106–115. DOI: https://doi.org/10.2307/1426315

Marsman, M., & **Wagenmakers, E.-J.** (2017). Three Insights from a Bayesian Interpretation of the One-Sided P Value. *Educational and Psychological Measurement, 77*(3), 529–539. DOI: https://doi.org/10.1177/0013164416669201

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & **Wagenmakers, E.-J.** (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123. DOI: https://doi.org/10.3758/s13423-015-0947-8

Oakes, M. W. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester; Wiley.

Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements. *Journal of the Royal Statistical Society. Series B (Methodological), 27*(2), 169–203.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: https://www.R-project.org/.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics, 12*(4), 1151–1172. DOI: https://doi.org/10.1214/aos/1176346785

Schervish, M. J. (1996). P Values: What They Are and What They Are Not. *The American Statistician, 50*(3), 203–206. DOI: https://doi.org/10.2307/2684655

Shao, J. (1998). *Mathematical Statistics*. New-York: Springer-Verlag. Retrieved from: https://www.springer.com/la/book/9780387953823.

Trafimow, D., & **Marks, M.** (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

van Ravenzwaaij, D., Cassey, P., & **Brown, S. D.** (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review, 25*, 143–154. DOI: https://doi.org/10.3758/s13423-016-1015-8

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & **Iverson, G. J.** (2008). Bayesian Versus Frequentist Inference. In: *Bayesian Evaluation of Informative Hypotheses*, 181–207. Springer, New York, NY. Retrieved from: https://link.springer.com/chapter/10.1007/978-0-387-09612-4_9. DOI: https://doi.org/10.1007/978-0-387-09612-4_9

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & **Wagenmakers, E.-J.** (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science, 6*(3), 291–298. DOI: https://doi.org/10.1177/1745691611406923

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.149.pr

How to cite this article: Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology,* 4(1): 31. DOI: https://doi.org/10.1525/collabra.149

Senior Editors: Simine Vazire, Victoria Savalei

Submitted: 23 February 2018 **Accepted:** 30 July 2018 **Published:** 24 August 2018

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS